Sains Malaysiana 43(10)(2014): 1599–1607

Imputing Missing Values in Modelling the PM_{10} Concentrations

(Mengganti Nilai Hilang dalam Pemodelan Kepekatan PM_{10})

NURADHIATHY ABD RAZAK, YONG ZULINA ZUBAIRI* & ROSSITA M. YUNUS

ABSTRACT

Missing values have always been a problem in analysis. Most exclude the missing values from the analyses which may lead to biased parameter estimates. Some imputations methods are considered in this paper in which simulation study is conducted to compare three methods of imputation namely mean substitution, hot deck and expectation maximization (EM) imputation. The EM imputation is found to be superior especially when the percentage of missing values is high as it constantly gives low RMSE as compared with other two methods. The EM imputation method is then applied to the PM_{10} concentrations data set for the southwest and northeast monsoons in Petaling Jaya and Seberang Perai, Malaysia which has missing values. Four types of distributions, namely the Weibull, lognormal, gamma and Gumbel distribution are considered to describe the PM_{10} concentrations. The Weibull distribution gives the best fit for the southwest monsoon data for Petaling Jaya. The lognormal distribution outperformed the others in describing the southwest monsoon in Seberang Perai. Meanwhile, for the northeast monsoon in both locations, gamma distribution is the best distribution to describe the data.

Keywords: Expectation maximization; mean imputation; missing value; PM_{10}; Weibull

ABSTRAK

Nilai hilang selalu menjadi masalah dalam analisis. Kebanyakan mengabaikan nilai hilang ini daripada analisis yang mungkin menyebabkan kepincangan dalam anggaran parameter. Beberapa kaedah gantian dipertimbangkan dalam kertas kerja ini dengan kaedah simulasi telah dijalankan untuk membandingkan kaedah-kaedah gantian tersebut iaitu penggantian menggunakan min, geladak panas dan jangkaan pemaksimuman (EM). Gantian EM didapati yang terbaik terutama apabila peratus nilai hilang adalah tinggi kerana ia berterusan memberi RMSE yang rendah berbanding dua kaedah yang lain. Kaedah gantian EM ini kemudiannya diaplikasikan pada set data kepekatan PM_{10} bagi monsun barat daya dan timur laut di Petaling Jaya dan Seberang Perai, Malaysia yang mempunyai nilai hilang. Empat jenis taburan, iaitu taburan Weibull, lognormal, gama dan Gumbel dipertimbangkan untuk menggambarkan kepekatan-kepekatan PM_{10}. Taburan Weibull memberi kesesuaian terbaik untuk data monsun barat daya bagi Petaling Jaya. Taburan lognormal pula mengatasi yang lain dalam menggambarkan monsun barat daya di Seberang Perai. Manakala bagi monsun timur laut di kedua-dua kawasan, taburan gama adalah taburan yang terbaik yang menggambarkan data tersebut.

Kata kunci: Jangkaan pemaksimuman; min gantian; nilai hilang; PM_{10}; Weibull

INTRODUCTION

Missing values are common phenomenon in almost all research studies, not to mention in the environmental and air pollution studies. The most common approach to handle missing values is by deleting those observations with incomplete information from the study. Such method is known as complete case analysis (Schafer 1997). However, Allison (2001) noted that this approach reduces the sample size and power of study. In addition, it may produce inefficient result, especially when the amount of missing values is large (Barzi & Woodward 2004) and data are not missing completely at random. Therefore, imputing the missing values is deemed necessary to avoid any misleading or devastating impact on the statistical inference due to the exclusion of subjects from the study in the analyses.

In air pollution studies, missing values may occur because of equipments malfunctioned or of errors in measurements (Noor & Zainudin 2008). In the literature, various techniques have been proposed to impute missing values in environmental data (Junninen et al. 2004; Norazian et al. 2008). For example, Fitri et al. (2010) and Noor et al. (2006) applied a simple method, namely mean top bottom technique to replace missing values in PM_{10} concentrations data set. However, Noor et al. (2006) found that this method performed well only when the number of missing values is small. Meanwhile, Shaadan et al. (2012) used nearest neighbour imputation method to impute incomplete PM_{10} concentration data in their study. In this paper, three methods namely the mean substitution, hot deck and expectation maximization (EM) imputation are considered. A simulation study is carried out to compare the performance of these methods.

The best method of imputing missing values is then applied to the PM_{10} concentrations data set which contains missing values. PM_{10} refers to the particles with diameters up to 10 micrometer which contains in the emission produced by motor vehicles, industrial activities and other natural sources. Jamal et al. (2004) suggested that the exposure to high PM_{10} level can cause acute morbidity such as respiratory diseases and cardiovascular diseases. In addition, it also increases mortality risks (Dominici et al. 2003). The PM_{10} concentrations data from two monitoring stations, Petaling Jaya and Seberang Perai are used in this study. The data is divided into two different monsoons namely the southwest and northeast monsoon. It is believed that the concentrations of PM_{10} in these two locations are different.

Recently, there have been a number of studies on the prediction of the exceedences and return period of the PM_{10} critical concentrations in the literature (Fitri et al. 2010; Noor et al. 2011). In most of these studies, several probability distributions were fitted to the observed PM_{10} concentrations data set. Then, the best fitting distribution was used to make better decision and prediction about the PM_{10} concentrations. Several types of probability distributions have been used to fit the PM_{10} concentrations such as the Weibull (Lu 2004), lognormal (Noor et al. 2011), gamma (Sansuddin et al. 2011) and Frechet and Gumbel distribution (Fitri et al. 2011). In this study, the Weibull, lognormal, gamma and Gumbel distribution are considered to describe the data set after imputing the missing values. It is expected that the monsoonal differences may affect the ambient PM_{10} concentrations level. The PM_{10} level is higher during the southwest monsoon (May – September) compared to the northeast monsoon (November – March) because of the dry weather condition (Fitri et al. 2010). The best fitting distribution is selected based on the performance indicators and the quantile-quantile (Q-Q) plot.

STUDY AREA

Petaling Jaya is located in the west coast of Malaysia with a geographical coordinate of 3^{o} 06' north latitude and 101^{o} 39' east longitude. The area covers 97.2 km^2 of the Petaling district. Petaling Jaya was the first satellite town developed to accommodate a high density of Kuala Lumpur population in 1950's. The town is packed with more than half a million population and a number of industrial areas. There are about 2200 industrial projects in Petaling Jaya. These industries consist of various sectors including chemical production, electronic and electrical industry, machine manufacturing and fabricated metal products (Majlis Bandaraya Petaling Jaya 2005). Petaling Jaya is located in the centre of Klang Valley, surrounded by other industrial and residence areas such as Kuala Lumpur, Subang and Shah Alam.

On the other hand, Seberang Perai is situated in the northern Peninsular of Malaysia covering a 738.41 km^2 area of Penang state. Geographically, latitude and longitude of Seberang Perai is 5° 21' north and 100° 24' east. The city is crammed with a large number of population and industries. The Population and Housing Census of Malaysia (2010) reported that there were 838999 residents of Seberang Perai in the year of 2010. In addition, more than 600 projects from various industrial sectors operate in Seberang Perai, such as the textile, electric and electronic as well as manufacturing industries. Rapid growth of industrial activities and traffic densities has affected the air quality in both areas. The emissions from such sources sometime have led to an unhealthy level of Air Pollution Index (API) which is measured on the basis of PM_{10} concentrations level. Besides, pollutants that spread from nearby areas have also affected the air quality surroundings and increased the level of PM_{10} concentrations in both areas.

MATERIALS AND METHODS

DATA

Data on PM_{10} concentrations was obtained from the website of Department of Environmental Malaysia. An average of the PM_{10} concentrations at 5.00 pm was chosen because it is expected that during this time, the ambient PM_{10} concentrations reach their high level resulting from the peak traffic density and industrials activities. Two monitoring stations located at two different areas, namely Petaling Jaya and Seberang Perai are considered. The data was grouped into two different monsoons, the southwest (May 2009 to September 2009) and northeast monsoon (November 2009 to March 2010) for each location. Data that has no information on PM_{10} concentrations level was considered missing.

MISSING DATA IMPUTATION METHOD

Missing data can be treated by either single or multiple imputation method. This study considers single imputation method where each missing item is imputed by only one estimated value. Meanwhile, multiple imputation is a method in which missing data are replaced with a set of plausible values (Clark et al. 2003). There are three methods of handling missing values considered in this study namely; mean substitution, hot deck and EM imputation.

Mean substitution is an imputation technique where missing items for any variable are filled in with the average of the observed value of that particular variable (Schafer & Graham 2002). Let x represents PM_{10} concentrations data with n observations for a particular monsoon and monitoring station. The data set contains missing and observed component denoted by x_{mis} and x_{obs}, respectively. Therefore, by mean substitution method, all r items in the x_{mis} were replaced by the average of the observed component, x_{obs}, that is

$$\bar{x}_{obs} = \frac{\sum_{i=1}^{n-r} x_{obs,i}}{n-r}.$$

In the hot deck imputation method, the imputed values were determined using the *k*-nearest neighbour (*k-nn*) method. The most similar example from the data set was determined by *k-nn* method and the missing values were imputed by the value found in such example. Given a data set with two variables, *x* and *y*, where there are some items in *y* are missing. Let y_j be a missing observation, the missing value y_j is substituted first by finding the difference between observed x_j and the nearest neighbours of x_j, x_i = (.., x_{j-1}, x_{j+1},..) using the following formula

$$D = \left| x_i - x_j \right|.$$

Then y_j is replaced with the observed y_k value for which the difference between x_k to x_j is the smallest.

Another method of imputing missing values is the EM imputation method. In this study, NORM package in the R software was used for the EM imputation. This program was developed to handle multivariate normal data with missing values using the theory of EM algorithm developed by Dempster et al. (1977). The algorithm consists of two iterative steps called the expectation step or the E-step and the maximization step or the M-step. Let *x* be a set of data that contains missing components, x_{mis} and observed components, x_{obs}. In any incomplete data problem, the density function can be written as

$$f\left(x|\theta\right) = f\left(x_{obs}|\theta\right) f\left(x_{mis}|x_{obs},\theta\right)$$

and the log-likelihood function is

$$\ell\left(\theta|x\right) = \ell\left(\theta|x_{obs}\right) + \log f\left(x_{mis}|x_{obs},\theta\right).$$

The second term of the above log-likelihood cannot be computed as x_{mis} component is unobserved. Therefore, the EM algorithm solves this problem by computing the conditional expectation $Q\left(\theta|\theta^{(m)}\right)$ of the complete-data log-likelihood function given the x_{obs} and current fit of θ,

$$Q\left(\theta|\theta^{(m)}\right) = E_{\theta(m)}\left(\ell_c\left(\theta;x\right)|x_{obs}\right).$$

Then, in the M-step, $Q\left(\theta|\theta^{(m)}\right)$ is maximized to obtained new parameter estimates, $\theta^{(m+1)}$. These steps are repeated until convergence

$$Q\left(\theta|\theta^{(m)}\right) \geq Q\left(\theta|\theta^{(m)}\right).$$

Then, random imputations of the missing data were drawn from multivariate normal distribution with the parameter values $\theta^{(m+1)}$. Although this method assumes multivariate normal data, this method can be applied to non-normal data by using suitable transformations to normality.

SIMULATION STUDY

A simulation study was conducted to examine how the parameters were affected by applying different methods of handling missing data. Data with two variables, *X* and *Y* were generated from normal distribution, with sample sizes of *n*, varying from 30 to 300. The mean (μ_X) and standard deviation (σ_X) of *X* are 50 and 10 respectively, while *Y*, the μ_Y=125 with σ_Y = 25. The correlation between both variables, ρ_{XY}, was set at 0.70. Only variable *Y* was made missing completely at random with different levels of missingness namely; 10, 30 and 50% respectively. As mentioned earlier, mean substitution, hot deck and EM imputation were considered in the analysis.

Three parameters were estimated namely; the mean of *Y* (μ_Y), standard deviation of *Y* (σ_Y) and the correlation coefficient (ρ_{XY}). The performance of these methods was measured based on the root mean squared error (RMSE). A smaller RMSE is desirable, since it indicates that a predicted value is closer to the exact value, therefore more accurate (Schafer & Graham 2002). The simulation process was repeated 10000 times for each combination of *n* and the percentage of missing values. The analysis was conducted using the R software**.** Figure 1 summarizes the steps involved in the simulation study.

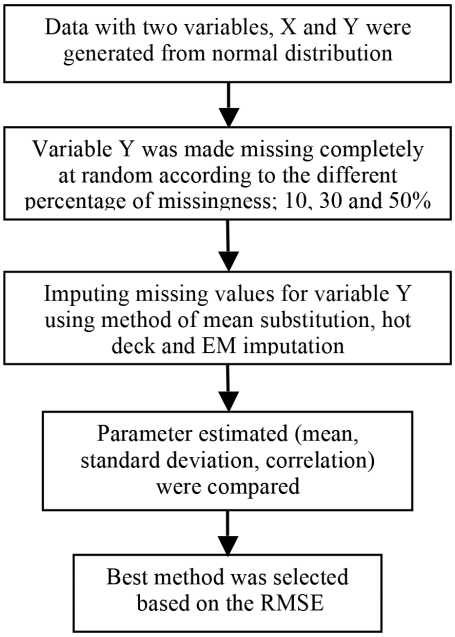

FIGURE 1. The flow chart of the simulation study

PROBABILITY DISTRIBUTION

Four types of probability distributions namely the Weibull, gamma, lognormal and Gumbel distribution were considered to fit the PM_{10} concentrations that had been imputed for the missing values. These are amongst the common statistical distribution used to fit environmental pollutions data and other meteorological data. The parameters were estimated by maximum likelihood estimation method.

The goodness of fit of these distributions was measured using four performance indicators; RMSE, coefficient of determination (R^2), mean absolute error (MAE) and Akaike Information Criterion (AIC). These measurements were used to determine the agreement between the predicted (p_i) and observed (x_i) concentrations of the PM_{10}. The predicted values were generated using the parameters obtained from the maximum likelihood estimation for all four distributions. The RMSE is given as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(x_i - p_i)^2}{n}}.$$

Meanwhile, the formula for the R^2 is

$$R^2 = \left(\frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(p_i - \overline{p})}{\sigma_x \sigma_p} \right)^2$$

where $\overline{x}$ and $\overline{p}$ are the mean of the observed and predicted data, respectively, σ_x and σ_p are the standard deviation of observed and predicted data, respectively. The formula for MAE is given as:

$$MAE = \frac{\sum_{i=1}^{n} |x_i - p_i|}{n}.$$

The AIC was calculated using the following formula

$$-2 \ln L + 2k,$$

where ln L is the logarithm of the likelihood function of the propose model and k is the number of parameters. A probability distribution that best fits the PM_{10} concentrations should have the smallest value of RMSE, MAE and AIC, and the largest value of R^2.

In addition, the quantile-quantile (Q-Q) plot was used to compare the distribution of PM_{10} data and the other four tested distributions graphically. The points lie approximately on a straight line $y = x$ suggests that two distributions being compared are similar. Figure 2 shows the steps involved in fitting the PM_{10} concentrations.

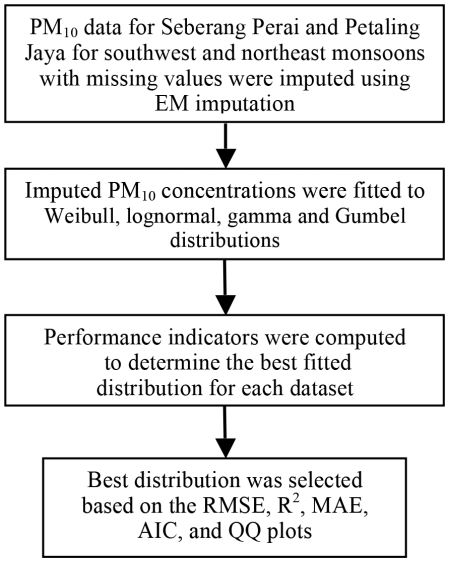

FIGURE 2. The flow chart of fitting PM_{10} concentrations data

RESULTS

Table 1 shows the descriptive statistics and the percentage of missing values available in the PM_{10} concentrations data set. There are about 20 to 45% of missing values identified. These missing values were imputed using the EM imputation method, since the simulation study suggested that this method yields better parameter estimations, compared with the other two methods.

Table 2 shows the RMSE of the estimation of the mean, standard deviation and coefficient correlation parameters for different methods of handling missing data. It can be seen that, as sample size increases, the RMSE values become lower. With a large sample sizes, the estimations seem closer to the true value. On the other hand, the higher the percentage of missing values, the higher is the RMSE. The estimations tend to deviate from the true value when there are too many missing values in the data.

TABLE 1. Descriptive statistics of PM_{10} concentrations

Parameter	Petaling Jaya		Seberang Perai	
	Southwest monsoon (n=153)	Northeast monsoon (n=151)	Southwest monsoon (n=153)	Northeast monsoon (n=151)
Observed data	116	96	120	84
% missing values	24.18	36.42	21.57	44.37
Mean	53.18	40.47	38.75	35.99
Median	54	39.5	37	36
Standard deviation	14.67	9.61	11.82	6.42
Minimum	19	21	20	23
Maximum	93	65	74	53

TABLE 2. The RMSE of the estimation of the mean, standard deviation, and correlation parameters for the simulated data

	Percentage of missing values								
	10%			30%			50%		
	n=30	*n*=100	*n*=300	*n*=30	*n*=100	*n*=300	*n*=30	*n*=100	*n*=300
Mean:									
Mean Subs.	4.790	2.650	1.509	5.512	2.987	1.715	6.426	3.511	2.073
Hot Deck	4.802	2.645	1.538	5.468	3.042	1.878	6.262	3.589	2.372
EM	4.709	2.613	1.492	5.476	2.841	1.612	6.066	3.096	1.806
Standard deviation:									
Mean Subs.	3.610	2.226	1.657	5.578	4.559	4.237	8.630	7.716	7.436
Hot Deck	3.524	1.886	1.119	4.141	2.272	1.404	5.074	2.817	1.793
EM	3.499	1.920	1.092	4.201	2.155	1.241	4.842	2.601	1.463
Correlation:									
Mean Subs.	0.114	0.069	0.049	0.176	0.135	0.121	0.260	0.222	0.211
Hot Deck	0.108	0.057	0.032	0.132	0.068	0.040	0.167	0.085	0.051
EM	0.106	0.059	0.032	0.136	0.063	0.036	0.158	0.079	0.042

It is also found that the RMSE for the mean parameter obtained using the mean substitution method is smaller than that of using the hot deck. On the contrary, for the standard deviation and correlation parameters, the RMSE obtained using the hot deck is smaller than that of using the mean substitution. The estimations by mean substitution deviate remarkably from the true value particularly when there are large amount of missing values. This shows that although the mean substitution method preserved the true mean value, the shape of the distribution distorted and the relationship of the variables are much affected. Overall, the RMSE of the estimation of all parameters by EM imputation are better than those obtained using the hot deck and mean substitution. The RMSE remains relatively small when the percentage of missing values reaches 50% regardless of the sample size.

Based on the simulation results, it can be concluded that the EM imputation method is preferable, compared with the other two even though the percentage of missing values is high. Therefore, this method was applied on the incomplete PM_{10} concentrations data set. In Table 3, the estimated values of the parameters of the Weibull, lognormal, gamma and Gumbel distributions of the imputed PM_{10} concentrations data are presented. The performance indicators were computed to find out which of the aforementioned distributions best fits the data and the results are given in Table 4.

Based on the results given in Table 4, the distribution that best fits the PM_{10} concentrations during the southwest monsoon in Petaling Jaya is the Weibull distribution, because the RMSE, MAE and AIC are the smallest compared with those of the other distributions. The R^2 between the predicted and observed concentrations for the Weibull distribution is the highest amongst others. A majority of the points lie on a straight line in the Q-Q plot of the Weibull distribution (Figure 3(a)).

As for the PM_{10} concentrations during the northeast monsoon in Petaling Jaya's sample data, the best fitting distribution for this sample data is the gamma distribution. The RMSE, MAE and AIC are the smallest and the R^2 is the largest when the sample data is fitted using the gamma distribution (Table 4). A strong linear trend is observed in the gamma distribution Q-Q plot (Figure 4(c)) suggesting good fit.

TABLE 3. Parameter estimates of the PM_{10} concentrations data

Distribution	Parameter		Petaling Jaya		Seberang Perai	
			Southwest monsoon	Northeast monsoon	Southwest monsoon	Northeast monsoon
Weibull	ξ	shape	4.1448	4.6786	3.2087	5.8901
	ϕ	scale	58.4936	44.5564	43.3350	39.3067
Lognormal	ξ	shape	3.9295	3.6805	3.6157	3.5823
	ϕ	rate	0.3065	0.2389	0.3026	0.1819
Gamma	ξ	location	11.8191	18.1249	10.9180	30.8214
	ϕ	scale	0.2225	0.4444	0.2803	0.8433
Gumbel	ξ	location	45.7651	36.1639	33.3864	33.3375
	ϕ	scale	14.3715	8.5434	9.5496	5.9723

TABLE 4. Performance indicators for the PM_{10} concentrations data

Distribution	Parameter	Petaling Jaya		Seberang Perai	
		Southwest monsoon	Northeast monsoon	Southwest monsoon	Northeast monsoon
Weibull	RMSE	2.1634	1.8852	3.6186	1.6120
	R^2	0.9729	0.9596	0.9221	0.9489
	MAE	1.6697	1.3975	2.5341	1.0239
	AIC	1246	1108	1202	1003
Lognormal	RMSE	4.9992	1.9572	2.2107	1.1956
	R^2	0.9145	0.9568	0.9653	0.9644
	MAE	3.4780	1.2569	1.3823	0.8384
	AIC	1271	1104	1171	992
Gamma	RMSE	3.5198	1.6078	2.4498	1.0895
	R^2	0.9446	0.9674	0.9567	0.9686
	MAE	2.6662	1.1283	1.5052	0.7680
	AIC	1259	1101	1176	991
Gumbel	RMSE	6.4619	2.9638	2.2360	2.0550
	R^2	0.8995	0.9318	0.9657	0.9357
	MAE	4.0760	1.6132	1.4095	1.2506
	AIC	1274	1109	1171	1000

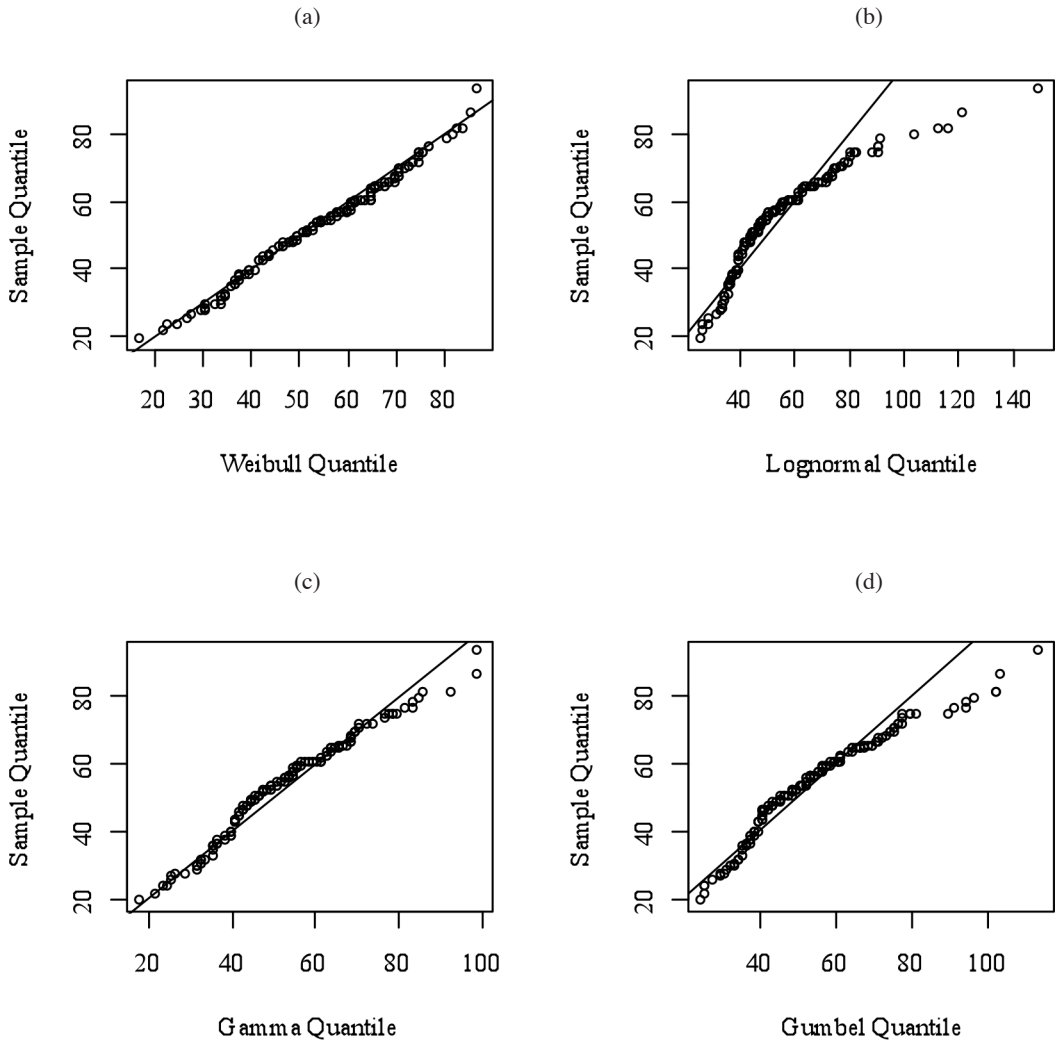

FIGURE 3. The Q-Q plot of the PM_{10} concentrations during the southwest monsoon in Petaling Jaya. a) Weibull, b) lognormal, c) gamma and d) Gumbel distributions

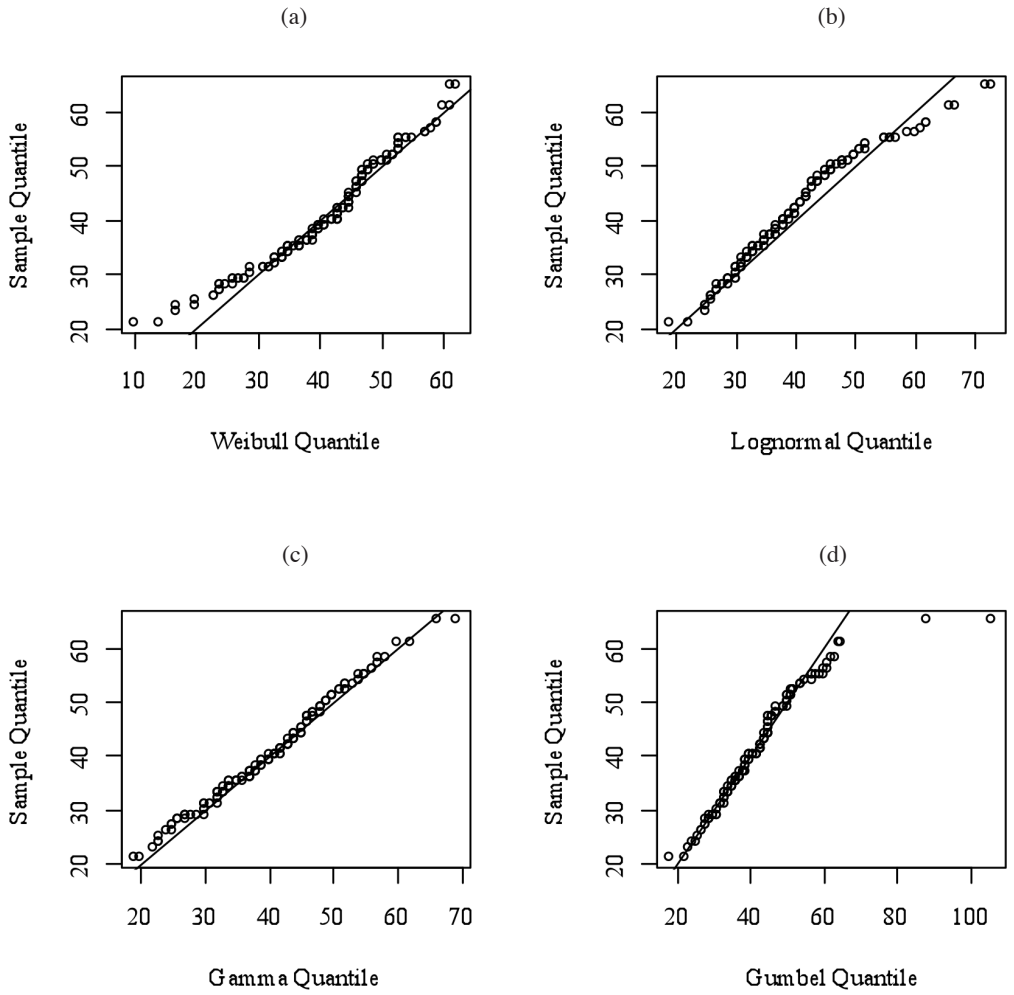

FIGURE 4. The Q-Q plot of the PM_{10} concentrations during the northeast monsoon in Petaling Jaya. a) Weibull, b) lognormal, c) gamma and d) Gumbel distributions

Unlike Petaling Jaya, the Weibull distribution turned out to be the worst distribution to describe the PM_{10} concentrations in Seberang Perai during the southwest monsoon (Table 4). Instead, the results showed that the lognormal distribution is the best. This is confirmed by the Q-Q plot (Figure 5(b)) of the lognormal distribution. Meanwhile, the RMSE, MAE, R^2 and AIC values in Table 4 show that the data for Seberang Perai during the northeast monsoon is best fitted using the gamma distribution. In Figure 6(c), a strong linear trend is observed in the Q-Q plot of the gamma distribution.

DISCUSSION

The simulation study showed that EM imputation was the best method in handling missing data compared with the mean substitution and hot deck method. The EM imputation method performed considerably well, even though the percentage of missing values is high. This can be seen by small values of RMSE. In addition, sample size and the number of missing values influence the estimation of the parameters. As sample size increases, the parameters estimated are closer to the true value. In contrast, the estimations tend to deviate from the true value as the percentage of missing values increases.

Therefore, the EM imputation method was applied to the PM_{10} concentrations data for two different locations, Petaling Jaya and Seberang Perai at different monsoons, to replace the missing values. There are about 20 to 45% missing observations in the data set. The imputed data sets were then fitted to four different probability distributions; the Weibull, lognormal, gamma and Gumbel. The parameters for each distribution were estimated by maximum likelihood estimation method. Based on the performance indicators and Q-Q plots, the best distribution was selected.

This study found that the gamma distribution was the most suitable distribution to represent the PM_{10} concentrations during the northeast monsoon in both locations. In Petaling Jaya, the Weibull distribution outperformed the other distributions while fitting the southwest monsoon data. Meanwhile, the lognormal distribution best fits the PM_{10} concentrations during the southwest monsoon in Seberang Perai. These can be seen

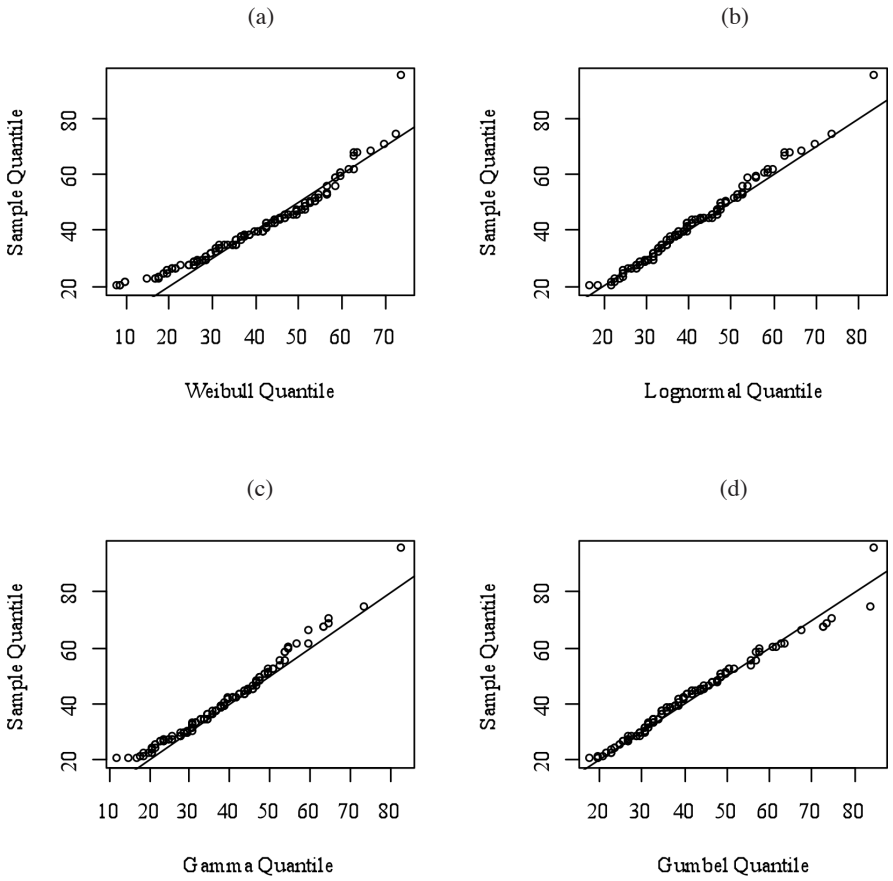

FIGURE 5. The Q-Q plot of the PM_{10} concentrations during the southwest monsoon in Seberang Perai. a) Weibull, b) lognormal, c) gamma and d) Gumbel distributions

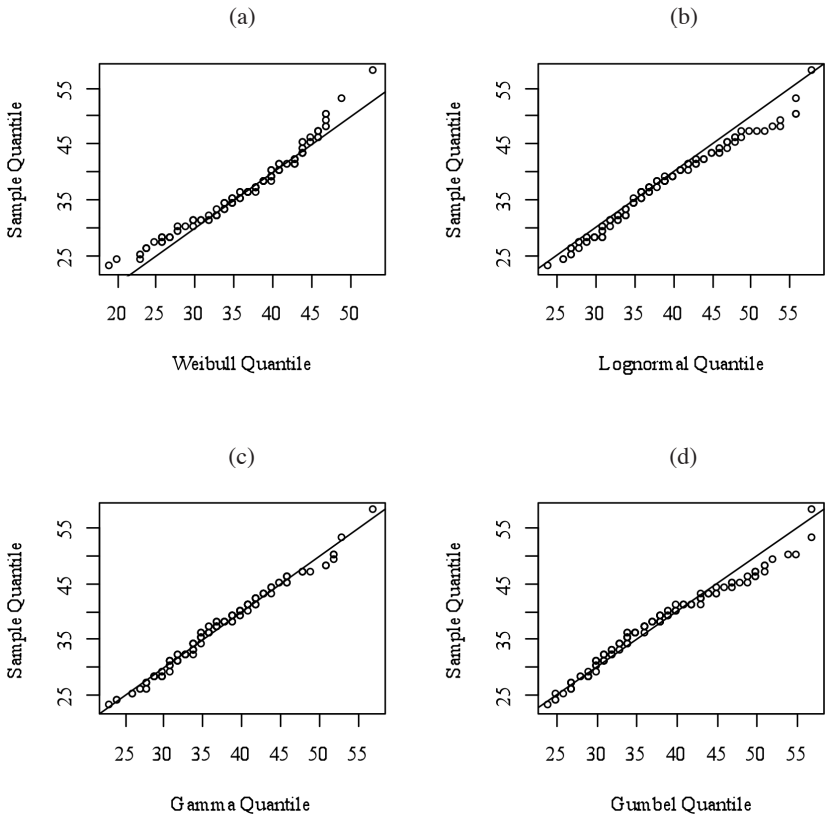

FIGURE 6. The Q-Q plot of the PM_{10} concentrations during the northeast monsoon in Seberang Perai. a) Weibull, b) lognormal, c) gamma and d) Gumbel distributions

from the measurements of the RMSE, MAE, R^2 and AIC where the lognormal gives the smallest error for Seberang Perai and the Weibull gives the smallest error for Petaling Jaya. The difference may be due to the mean concentrations of PM_{10} in Petaling Jaya was higher than Seberang Perai during this monsoon. Fitri et al. (2010) claimed that the Weibull distribution is better to describe the high PM_{10} concentrations.

The contribution of the current study is that once the probability distribution is identified, the behaviour of the distribution of the PM_{10} can be further understood, such as the expected level, exceedences and return period of the PM_{10}. In addition, this finding may help to model the effect of the PM_{10} concentrations with the presence of other factors.

ACKNOWLEDGEMENTS

We are most grateful to the Department of Environment and Department of Meteorological Malaysia for the data. In addition, thanks to University of Malaya and Ministry of Education (MOE), Malaysia for the financial support. We also wish to thank the referees for their helpful comments and suggestions.

REFERENCES

Allison, P.D. 2001. *Missing Data*. California: Thousand Oaks, Sage.

Barzi, F. & Woodward, M. 2004. Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology* 160: 34-45.

Clark, T.G., Bradburn, M.J., Love, S.B. & Altman, D.G. 2003. Survival Analysis Part IV: Further concepts and methods in survival analysis. *British Journal of Cancer* 89: 781-786.

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B(Methodological)* 39: 1-38.

Department of Statistics. 2011. Population distribution and basic demographic characteristics 2010. http://www.statistics.gov.my/portal/. Assessed on 29 November 2011.

Dominici, F., McDermott, A., Zeger, S.L. & Samet, J.M. 2003. National maps of the effects of particulate matter on mortality: Exploring geographical variation. *Environmental Health Perspectives* 111: 39-43.

Fitri, M.D.N.F., Ramli, N.A. & Yahaya, A.S. 2011. Extreme value distribution for prediction of future PM_{10} exceedences. *International Journal of Environmental Protection* 1: 28-36.

Fitri, M.D.N.F., Ramli, N.A., Yahaya, A.S., Sansuddin, N., Ghazali, N.A. & Al Madhoun, W. 2010. Monsoonal differences and probability distribution of PM_{10} concentration. *Environmental Monitoring Assessment* 163: 655-667.

Jamal, H.H., Pillay, M.S., Zailina, H., Shamsul, B.S., Sinha, K., Zaman Huri, Z., Khew, S.L., Mazrura, S., Ambu, S., Rahimah, A. & Ruzita, M.S. 2004. *A Study of Health Impact & Risk Assessment of Urban Air Pollution in Klang Valley, Malaysia*. Kuala Lumpur: UKM Pakarunding Sdn Bhd.

Junninen, H., Niska, H., Tuppurrainen, K., Ruuskanen, J. & Kolehmainen, M. 2004. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment* 38: 2895-2907.

Lu, H.C. 2004. Estimating the emission source reduction of PM_{10} in central Taiwan. *Chemosphere* 54: 805-814.

Majlis Perbandaran Petaling Jaya. 2005. *Maklumat Asas Petaling Jaya*. Petaling Jaya: Majlis Perbandaran Petaling Jaya.

Norazian, M.N., Shukri, Y.A., Azam, R.N. & Mustafa Al Bakri, A.M. 2008. Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia* 34: 341-345.

Noor, N.M., Tan, C.Y., Abdullah, M.M.A., Ramli, N.A. & Yahaya, A.S. 2011. Modelling of PM_{10} concentration in industrialized area in Malaysia: A case study in Nilai. *2011 International Conference on Environment and Industrial Innovation IPCBEE,* Vol.12. Singapore: IACSIT Press.

Noor, N.M. & Zainudin, M.L. 2008. A review: Missing values in environmental data sets. In *Proceeding of International Conference on Environment*.

Noor, N.M., Yahaya, A.S., Ramli, N.A. & Abdullah, M.M.A. 2006. The replacement of missing values of continuous air pollution monitoring data using mean top bottom imputation technique. *Journal of Engineering Research & Education* 3: 96-105.

Sansuddin, N., Ramli, N.A., Yahaya, A.S., Fitri, M.D.N.F., Ghazali, N.A. & Al Madhoun, W.A. 2011. Statistical analysis of PM_{10} concentrations at different locations in Malaysia. *Environmental Monitoring Assessment* 180: 573-588.

Schafer, J.L. & Graham, J.W. 2002. Missing data: Our view of the state of the art. *Psychological Methods* 7: 147-177.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Shaadan, N., Deni, S.M. & Jemain, A.A. 2012. Assessing and comparing PM_{10} pollutant behaviour using functional data approach. *Sains Malaysiana* 41(11): 1335-1344.

Nuradhiathy Abd Razak
Institute of Graduate Studies, University of Malaya
50603 Kuala Lumpur
Malaysia

Yong Zulina Zubairi*
Centre for Foundation Studies in Science
University of Malaya
50603 Kuala Lumpur
Malaysia

Rossita M. Yunus
Institute of Mathematical Sciences
University of Malaya
50603 Kuala Lumpur
Malaysia

*Corresponding author; email: yzulina@um.edu.my

Received: 30 July 2013
Accepted: 13 February 2014